

Abstract Title Page

Not included in page count.

Title:

What is Gained from a Randomized Controlled Trial and What is Lost? Planning, Policy, and Fear: Effects and Lessons Learned from a Cluster-Randomized Trial Assessing a Community-Advocacy Program's Professional Development Initiative for Early Childhood Educators

Author(s):

Kim Sprague, Brown University, Fred Glantz, Kokopelli Associates LLC, Pamela Raya-Carlton, RMC ResearchCorporation, Diane Schilder, Evaluation-Analysis Solutions, Inc.

Kim Sprague, The Education Alliance at Brown University, kim_sprague@brown.edu

Fred Glantz, Kokopelli Associates LLC, fred@kokopelliassociates.com

Pamela Raya-Carlton, RMC ResearchCorporation, prayacarlton@rmccorp.com

Diane Schilder, Evaluation Analysis Solutions, Inc. dschilder@eval-inc.com

Abstract Body

Background/context:

Description of prior research and/or its intellectual context and/or its policy context.

Extensive research has been conducted to explore the reasons why children are at risk of early reading failure. More specifically, research has sought to identify ways in which Early Childhood Educators (ECEs) can promote children's emergent language and literacy skills. Such research has indicated children at risk perform better when supported by high quality literacy environments and ECEs who appropriately promote reading skills (Dickinson, & Sprague, 2001; Dickinson, & Smith, 2001; National Early Literacy Panel, 2006; Snow, Burns, & Griffin, 1998). Findings reflect the urgent need to identify *effective* scientifically-based programs to improve children's early reading skills, especially for children at risk of early reading failure. As a result, rigorous and scientifically-based research must be conducted to assess the ever-growing number of early childhood education and professional development programs (Welch-Ross, Wolf, Moorehouse, & Rathgeb, 2006).

Purpose/objective/research question/focus of study:

Description of what the research focused on and why.

Rigorous research provides information that will allow ECE programs to select interventions that have a scientifically based track record of effectiveness in increasing teachers' skills and teaching quality. This paper shares implementation and impact results as well as the lessons learned in conducting an evaluation of an intervention developed to increase Early Childhood Educators (ECEs) knowledge and skills and ultimately improvements in children's language and readiness skills as a result. Evaluators developed the rigorous design and external evaluation plans to assess the effectiveness of this community advocacy-service delivery organization's professional development program. The evaluation was organized as two studies: (1) an **Impact Study** designed to estimate the effects of participating in the professional development program on early education providers as well as the children in their care; and (2) an **Implementation Study** designed to provide information about the way in which the professional development interventions were implemented (i.e., fidelity of implementation or fidelity to plans/model).

The evaluation plan employs a cluster-randomized trial (CRT) at the classroom level, to assess the effectiveness of the program in improving ECE practice as well as children's early reading skills. In addition, the consideration of implementation context in the interpretation of the results and what should be done when planning to implement will also be discussed. As a result evaluators hope to provide important information to those planning similar studies in such programs and other complex social settings. The paper will also consider the following: What methods should be employed in a randomized controlled trial (RCT) of an intervention delivered by a community advocacy program? What is required in terms of policy for implementation and collaboration to conduct such efforts? What is required to appropriately assess implementation?

Setting:

Specific description of where the research took place.

The location of sites is within a Northeastern state in a small urban center and its outlying areas. The participating Head Start program was recruited by the intervention program to include all sites and classes across the local service area.

Population/Participants/Subjects:

Description of participants in the study: who (or what) how many, key features (or characteristics).

Participants were local ECEs in a Head Start Program serving low-income children and their families from high poverty, culturally diverse communities. There were a total of 84 classrooms in the original proposed three-year study: 20 classrooms were to be included in the fidelity cohort and the remaining 64 classrooms in the main cohort. Refer to Figures 1-4 in Appendix B for more information on participants included in the fidelity cohort and the originally planned three-year design (sample, data collection, etc.).

Intervention/Program/Practice:

Specific description of the intervention, including what it was, how it was administered, and its duration.

The organization's program model had three primary objectives: (1) provide a research-based professional development intervention to ECEs with a level of depth, dosage, and duration needed to transform teaching practice in the areas of literacy and curriculum and assessment; (2) change the school readiness status and developmental trajectory of children as evidenced by their performance on formal school readiness assessments taken during the kindergarten year; and (3) close ethnic and racial gaps in school readiness by increasing the capacity of culturally and linguistically isolated ECEs who serve children from high need communities. The intervention combined elements of other successful programs and curriculum such as the *HeadsUp! Reading* known as the HUR program (The National Head Start Association, & The Council for Professional Recognition, 2004) and the *Opening the World of Learning* or OWL program and curriculum (Schickedanz & Dickinson, 2004). These elements were used to create a local research-based early language and literacy professional development intervention primarily consisting of two courses: *An established training program and curriculum course; and a course developed to support curriculum implementation. In addition, mentoring support and other program-sponsored opportunities were provided (e.g., leadership training, clubs for peer-collaboration).* The HUR course component was to be implemented followed by the Early Literacy Curriculum course (ELC) which was based on the OWL program and curriculum. While each course was provided over a shorter period of time, the content was to remain unchanged. Participants in the fidelity cohort were to receive 37.5 hours of HUR, 37.5 hours of ELC, and 30 hours of program mentoring.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

The evaluation design as proposed included two successive components: the first involved the assessment of implementation fidelity for all interventions delivered (i.e., *was the intervention delivered as intended?*); the second involved the assessment of impacts (i.e., *did the intervention improve student outcomes?*). The community-advocacy organization worked with evaluators to apply for a professional development grant offered by the US Department of Education. The grant gave competitive preference priority to rigorous evaluation designs assessing program

effectiveness. The evaluation plan employed a cluster-randomized trial (CRT) at the classroom level, to assess the effectiveness of the intervention (Raudenbush, Spybrook, & Congdon, 2004). For the first and second year of implementation, half of the center-based classrooms were to be assigned, at random, to the treatment group and half to the control (delayed-treatment) group. An equal proportion of home-based classrooms were to be assigned at random to the treatment and the control (delayed-treatment) group in the second year of implementation. Upon receipt of the award, design and evaluation plans were further developed by evaluators in consultation with the organization as details/specifications of their intervention were developed to ensure it would meet grant requirements and enable their group to conduct the work as proposed. To assess the impact of the intervention, evaluators collected data from participating classes/teachers and children and analyzed these data in concert with secondary program and participant data provided via the organization (now referred to as the program).

Fidelity Cohort Rationale. First, the fidelity cohort provided the program and evaluators with the opportunity to establish processes for maintaining implementation fidelity and to develop the necessary measures with the required validity and reliability to consistently evaluate the model. Second, it provided preliminary information on the success of implementation as well as the effectiveness of the program intervention in time for inclusion in a first-year report to Congress, as required. This cohort was compressed into a single semester given the timing of the award and the requirement for Institutional Review Board (IRB) at Brown University review and approval. After the design was approved by the federal program officer the plans were then submitted for IRB approval. The control group or delayed-treatment group participating in fidelity period was to be “rolled-in” to participate in the intervention. As a result of the fidelity period, the rigor of the evaluation was retained via a post-test only design for the first course (valid given the RCT component), and a pretest and post-test design for the second course (valid given evaluators caveats regarding time of test-retest) as implemented (Shadish, Cook, & Campbell, 2002). Refer to Figures 3 and 4 in Appendix B for more information about the originally planned three-year design and the first component, the fidelity period. Although adequate power for analysis was estimated for the entire sample, preliminary findings were to be included in the report to Congress so the fidelity cohort needed adequate statistical power to detect meaningful effects. An $MDE=0.40-0.60\sigma$ is generally considered to be a medium-size effect in similar evaluations (Bloom, Richburg-Hayes, & Rebeck-Black, 2005).

Data Collection and Analysis:

Description of plan for collecting and analyzing data, including description of data.

Evaluators collected primary data from participating classes/teachers and children and analyzed these data in concert with available secondary data provided by the program. Data for this research study were collected at two time periods during the abbreviated period of implementation – the fidelity cohort. The initial data collection took place immediately following IRB approval which coincided with the completion of the program implementation of the HUR course (the first half of the intervention). Therefore, no opportunity was provided for the collection of pretest data to establish a baseline for the participants and the children they serve because program implementation began prior to the receipt of IRB approval. However, the delay in IRB approval resulted in the benefit of additional analysis possibilities in which to examine treatment contrasts. Because the initial data collection occurred as the HUR course was completed and the design is a randomized control trial (RCT), the analysis of implementation results followed each discrete half of the intervention.

Implementation: Value-Added Design Modifications. The first or initial assessment was conducted at the end of the implementation of HUR and prior to the next, ELC. Therefore, the initial assessment was a **post-test** for this component of the intervention. Because participants were randomly assigned, an assumption was that the treatment and control group participants would be equivalent at baseline. That assumption was tested. In addition, the initial assessment was a **pretest** for the second course-component of the ECEPD intervention, the ELC course. Therefore, data collected via the final test administered at the end of the implementation – fidelity cohort – period were used to evaluate the **added-value** for participants in the second course component in relationship to any impacts observed following the first course component. The results from the final test administration were also compared to those from the initial test administration to determine the impact the intervention had on participants receiving the entire intervention.

Measures. ECE observational measures included the ELLCO to assess classroom environment and practices (Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002). Both were supplemented with key practice components during the observation protocol including the quality and intensity of teacher-child interactions (Arnett, 1989). ECE interviews were developed using items taken from previous interviews/surveys to obtain information about their backgrounds, participation, and perspectives. Child measures included the Peabody Picture Vocabulary Test (PPVT-III) to measure receptive vocabulary and the Phonological Awareness Literacy Screening (PALS Pre-K) to measure literacy skills (Dunn, & Dunn, 1997; Ivernizzi, Sullivan, & Meier, 2001). At the end of each semester of the intervention period, the program was to provide the evaluators with secondary data collected through their registration and tracking systems to document participation levels (e.g., attendance, hours) as well as participant demographic information.

Analysis. Aggregate change in child achievement status between the two groups will be assessed. To account for the hierarchical nature of these data, the most appropriate analytic approach is multilevel modeling using HLM (Raudenbush & Bryk, 2002; Raudenbush, Spybrook, Liu, & Congdon, 2004). Also using HLM, impact will modeled as a function of variables at both levels. For the first year sample, the estimate of the mean difference between the intervention and control groups on the outcome measures will be obtained by simply including classroom group assignment (X_j) as a predictor at level-2; the coefficient γ_{01} directly interpreted as the impact of the intervention on the outcome. Other covariates (e.g., aggregate pretest scores, child demographic characteristics) included take into account a variety of child and classroom characteristics known to be related to early language development, pre-reading skills, and reading comprehension. Teacher differences were individually assessed though the sample was small.

Findings/Results:

Description of main findings with specific details.

First year fidelity cohort results were mixed for both implementation and impacts. As proposed, to assess fidelity of implementation the Alliance relied heavily on data provided by the program. The analysis of fidelity period data provided by the program and coupled with qualitative data collected by evaluators reveals that the HUR course was implemented as planned (this was a course the program had a great deal of experience conducting). Data provided by the program reveal that their staff followed the course syllabi and the self-report data provided by the program indicated limited variation occurred in the implementation of this course. However, the

data provided by the program regarding mentoring and the curriculum course were insufficient to adequately answer questions about why gains were not sustained following the first half of the course implementation. The post hoc program data collected provided a limited picture of the curriculum course content yet revealed variability in the quality of implementation and depth of the content as well as in the nature of the mentoring visits. Since detailed plans were not in place prior to the implementation of the mentoring or the curriculum course and limited post hoc data were provided, the external evaluators were unable to assess the fidelity of these important aspects of the intervention. Data were also not provided to fully characterize the counterfactual, that is, to describe what happened in the absence of treatment. Several implementation issues were faced by the program including: (1) a loss of staff expertise in early language and literacy to advise curriculum implementation and training needed for the second course; (2) development was ongoing for the second course but development should have only occurred for tools and methods for tracking implementation; and, (3) program difficulties in developing appropriate tools for tracking implementation and collecting these data.

The first year of the intervention resulted in statistically significant differences for participants, as compared to non-participants, among key outcome scores *following HUR implementation* (the first of two professional development course components implemented). However, after participating in the second course component of the intervention, the ELC course focused on language and literacy instruction, the gains observed in the first half of the course were not substantially increased. Based on data collected via valid and reliable measures of classroom environment and quality, teachers in the treatment group scored higher than teachers in the control group (at Time 1 as well as Time 2). The sample was small and effect sizes were small. Refer to Figure 5 in Appendix B. Finally, children in the treatment group teachers' classes scored significantly higher overall than the children in the control group teachers' classes on outcome measures: an assessment of language skills (RCT post-test only design as appropriate given time constraints); and, letter identification (RCT pre-and post-test design appropriate given research on development). Effect sizes were modest but demonstrable. Refer to Figures 6-10 in Appendix B.

Conclusions:

Description of conclusions and recommendations of author(s) based on findings and over study. (To support the theme of 2009 conference, authors are asked to describe how their conclusions and recommendations might inform one or more of the above noted decisions—curriculum, teaching and teaching quality, school organization, and education policy.)

These studies are considered critical by the Department of Education to contribute to the field by addressing the primary research question of what interventions are effective, for whom and under what conditions. These studies should be considered in policy decisions regarding how to allocate finite resources to improve programs (Orr, 1999). Studies such as the one presented in this paper would provide evidence to select programs which increase the knowledge, skills, and ultimately the practice and quality of Early Childhood Education staff. In addition to the results presented and the potential to inform decisions, this study and others provide opportunities to consider the methods which should be employed in such RCT trials (e.g. what is required in terms of policy for implementation and collaboration to conduct such efforts). This paper seeks to understand and add to existing knowledge about both benefits of an experimental research design and the potential pitfalls in implementing such a design. Finally, the consequences a program shift in the purpose and conduct of the evaluation will be presented.

Appendixes

Not included in page count.

Appendix A. References

References are to be in APA format. (See APA style examples at the end of the document.)

- Arnett, J. (1989). "Caregivers in Day-Care Centers: Does Training Matter?" *Journal of Applied Developmental Psychology*, 5, 541-552.
- Bloom, H.S., Richburg-Hayes, L., & Rebeck-Black, A. (2005, November). *Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions*. MDRC Working Papers on Research Methodology. Manpower Demonstration Research Corporation New York, N.Y. Last retrieved December 1, 2005 from: <http://www.mdrc.org>.
- Dickinson, D. K. & Smith, M. W. (2001). Supporting language and literacy development in the preschool classroom. In D. K. Dickinson & P. O. Tabors, (Eds.), *Beginning literacy with language* (pp. 139-47). Baltimore, MD: Brookes Publishing Co.
- Dickinson, D.K. & Sprague, K. (2001). Chapter 18: The Nature and Impact of Early Childhood Care Environments on the Language and Early Literacy Development of Children from Low-Income Families. In Neuman and Dickinson, Neuman, S.B. & Dickinson, D.K., Eds. (2001). *Handbook of Early Literacy Research*. New York: The Guilford Press.
- Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (Rev. ed.) Circle Pines, MN: American Guidance Service.
- Invernizzi, M., Sullivan, A., & Meier, J. D. (2001). *Phonological Awareness Literary Screening: Pre-kindergarten*. Charlottesville, VA: University of Virginia.
- National Head Start Association & the Council for Professional Recognition. (n.d.). HeadsUp! Network, a service of National Head Start Association. Updated/retrieved July 2006. Last retrieved from: <http://www.heads-up.org>.
- The National Early Literacy Panel (2006). *Findings from the National Early Literacy Panel: Providing a Focus for Early Language and Literacy Development*. NCFL Annual Conference Louisville KY, March 19, 2006. Retrieved July 2006, from National Center for Family Literacy Web site: <http://www.famlit.org/loader.cfm?url=/commonspot/security/getfile.cfm&PageID=1968>
- Orr L. (1999). *Social experiments. Evaluating public programs with experimental methods*. Thousand Oaks, CA: Sage.

- Raudenbush, S. W., & Byrk, A. S. (2002) *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2004). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software*. Last retrieved January, 2006 from http://sitemaker.umich.edu/group-based/files/od_documentation/chapters1-8_vmarch09-05.pdf.
- Ricciuti, A.E., R.G. St.Pierre, W. Lee, A. Parsad & T. Rimdzius (2004). *Third National Even Start Evaluation: Follow-Up Findings From the Experimental Design Study*. US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Washington, D.C.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin Company.
- Schickendanz, J. & Dickinson, D. (2004). *Opening the World of Learning – OWL curriculum*. Developed in collaboration with Charlotte Mecklenburg Schools. New York: [Pearson Education Inc.](#), publishing as Pearson Learning Group. Copyright © 2004. All rights reserved.
- Smith, M.W., Dickinson, D.K., Sangeorge, A., & Anastasopoulos, L. (2002). *The Early Language and Literacy Classroom Observation*. Baltimore, MD: Brookes Publishing.
- Snow, C.E., Burns, S.M., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children*. Washington, DC: National Academy Press.
- Welch-Ross, M., Wolf, A. Moorehouse, M., & Rathgeb, C. (2006). Improving connections between professional development research and early childhood policies. In M. Zaslow & Martinez-Beck, I. (Eds.). *Critical Issues in Early Childhood Professional Development*. Baltimore, MD: Paul H. Brookes Publishing.
- What Works Clearinghouse (2005). *Key items to get right when conducting a randomized controlled trial in education*. Department of Education, Institute for Education Sciences. Last retrieved May 23, 2006 from http://www.whatworkshelpdesk.ed.gov/guide_RCT.pdf.

Appendix B. Tables and Figures
Not included in page count.

Figure 1. Fidelity cohort children

FIDELITY – PRETEST	COUNTS
Total population of children	420 children
Consented population	335 children (80%)
Consented (of the 220 original random selection)	173 of the original 220 (79%)
Completed (of original with consent)	157 children (91%)
Completed sample	228 children of the 220 target (100%)
FIDELITY – POSTTEST	COUNTS
Completed sample post-tested of those pretested	221 children (97%)

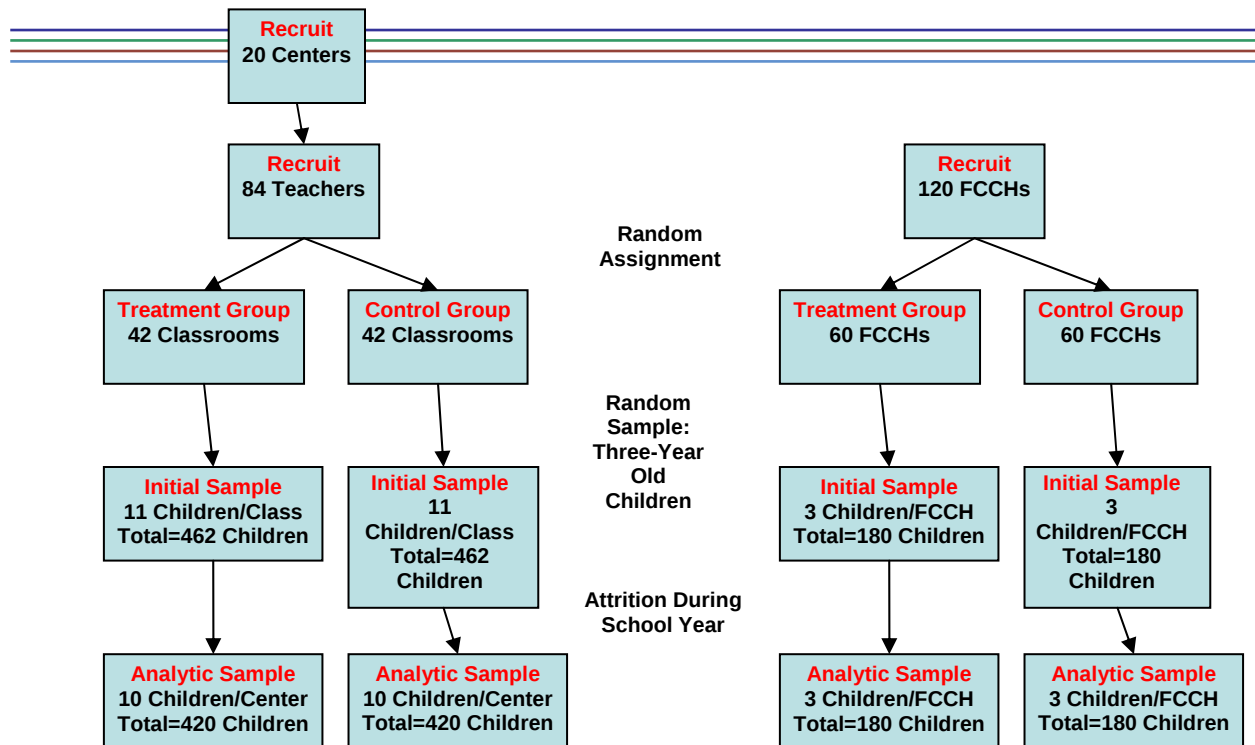
^a Total number of possible children across centers based on enrollment information provided

^b Total number of consents returned after blanketing entire center/classroom population

^c The additional 21% were then selected at random from the alternate list of those selected

^d All completed child cases have parental consent. The final data file contains two cases with no response or refusals for assessment.

Figure 2. Design sample summary (original three-year design)



Note: FCCH is Family Child Care Homes (originally proposed)

Figure 3. Teacher sample and observation/interview data collection timeline

		TEACHER GROUP	YEAR 1		YEAR 2		YEAR 3	
			Semester 1 Feb – June 07	Semester 2 Sept – Dec 07	Semester 3 Feb – June 08	Semester 4 Sept - Dec 08	Semester 5 Feb – June 09	
CENTER-BASED CLASSES	FIDELITY COHORT	Treatment <i>n</i> = 10	Receive PD Pre	Receive PD Post		Post		Post
		Control (delayed trt) <i>n</i> = 10	No PD Pre	No PD Pre	Receive PD Pre	Receive PD Post		Post
	MAIN COHORT	Treatment <i>n</i> = 32			Receive PD Pre	Receive PD Post		Post
		Control (delayed trt) <i>n</i> = 32			No PD Pre	No PD Pre	Receive PD Pre	Receive PD Post
HOME-BASED CLASSES	MAIN COHORT	Treatment <i>n</i> = 60			Receive PD Pre	Receive PD Post		Post
		Control (delayed trt) <i>n</i> = 60			No PD Pre	No PD Pre	Receive PD Pre	Receive PD Post

Figure 4. Child sample and assessment data collection timeline

		CHILD GROUP	YEAR 1		YEAR 2		YEAR 3	
			Semester 1 Feb – June 07		Semester 2 Sept – Dec 07	Semester 3 Feb – June 08	Semester 4 Sept - Dec 08	Semester 5 Feb – June 09
CENTER-BASED CLASSES	FIDELITY COHORT	Treatment n = 100	pre	post		Post (school data)		post (school data)
		Control (delayed trt) n = 100	1 st pre	2 nd pre	3 rd pre	Post		post (school data)
	MAIN COHORT	Treatment n = 320			Pre	Post		post (school data)
		Control (delayed trt) n = 320			1 st pre	2 nd pre	3 rd pre	post
HOME-BASED CLASSES	MAIN COHORT	Treatment n = 180			Pre	Post		post (school data)
		Control (delayed trt) n = 180			1 st pre	2 nd pre	3 rd pre	post

Note: Child sample size estimates include some anticipated within-year attrition.

Figure 5. Fidelity cohort data: classroom/teacher observations

		CONTROL MEAN (SD)	TREATMENT MEAN (SD)	Effect Size Significance
		n = 10	n = 10	
ELLCO <i>Literacy Environment Checklist</i> ^a	Time 1	20.30 (6.77)	27.30 (5.74)	.56 * $p = .02$
	Time 2	19.60 (4.88)	23.20 (5.53)	.35
ELLCO <i>Classroom Observation Total</i> ^b	Time 1	37.00 (9.42)	44.80 (8.70)	.43 ~ $p = .07$
	Time 2	41.70 (4.45)	42.90 (11.03)	.08
ELLCO <i>General Classroom Environment Subscale</i> ^a	Time 1	15.50 (3.69)	18.20 (3.68)	.37 $p = .12$
	Time 2	16.90 (2.64)	17.00 (3.80)	.02
ELLCO <i>Language, Literacy, Curriculum Subscale</i> ^a	Time 1	19.10 (5.45)	24.10 (5.26)	.47 * $p = .05$
	Time 2	21.80 (2.20)	23.60 (7.11)	.19
Early Literacy Activity Rating Scale	Time 1	4.90 (2.08)	8.10 (1.60)	.87 *** $p = .001$
	Time 2	6.00 (2.40)	6.40 (2.32)	.08
ECE Literacy Observation Checklist	Time 1	26.40 (6.65)	30.30 (5.68)	.32
	Time 2	25.70 (5.81)	32.20 (6.11)	.55 * $p = .03$

^a One of two subtests of the ELLCO Classroom Observation.

^b Score is a combination of others listed.

Figure 6. Impacts on Children's Letter Recognition: Post-Course 1

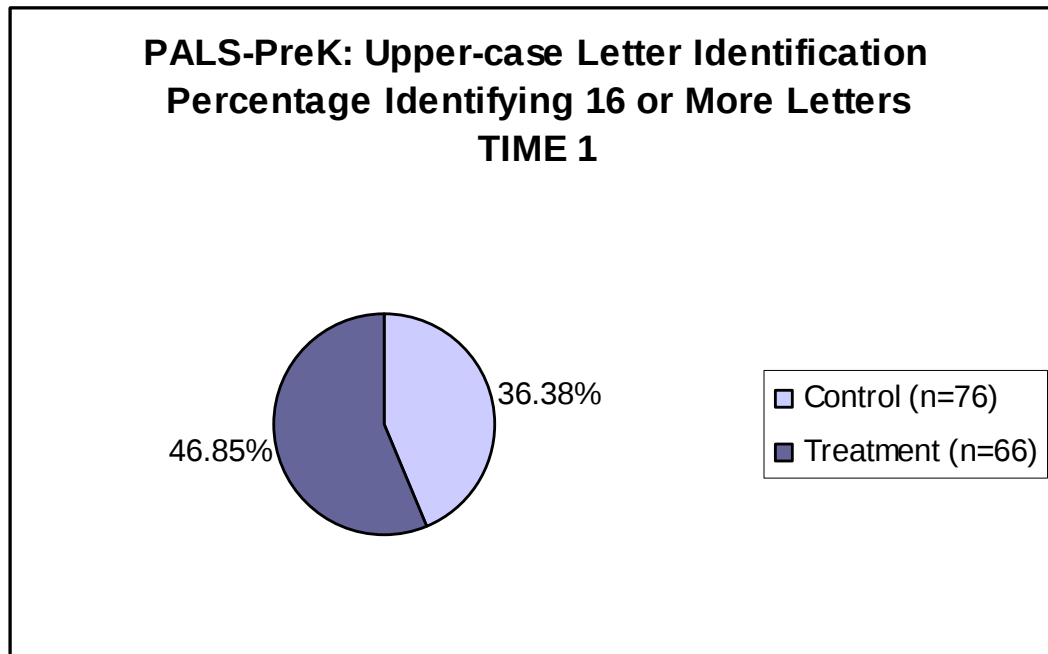


Figure 7. Impacts on Children's Letter Recognition: Post-Course 2

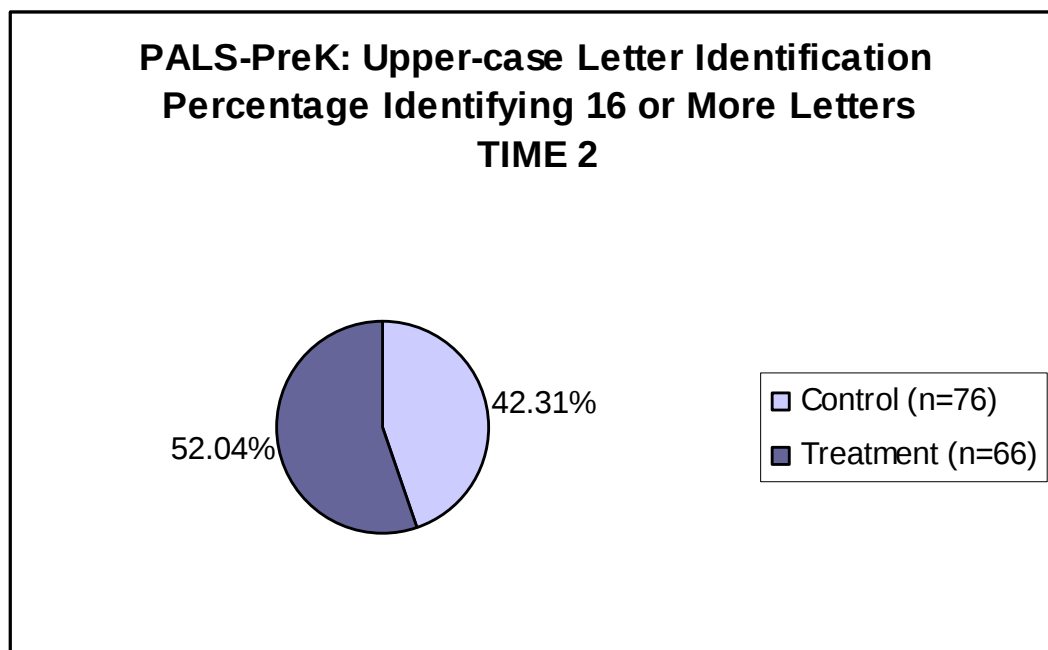


Figure 8. Impacts on Children's Receptive Vocabulary: Post-Course 1

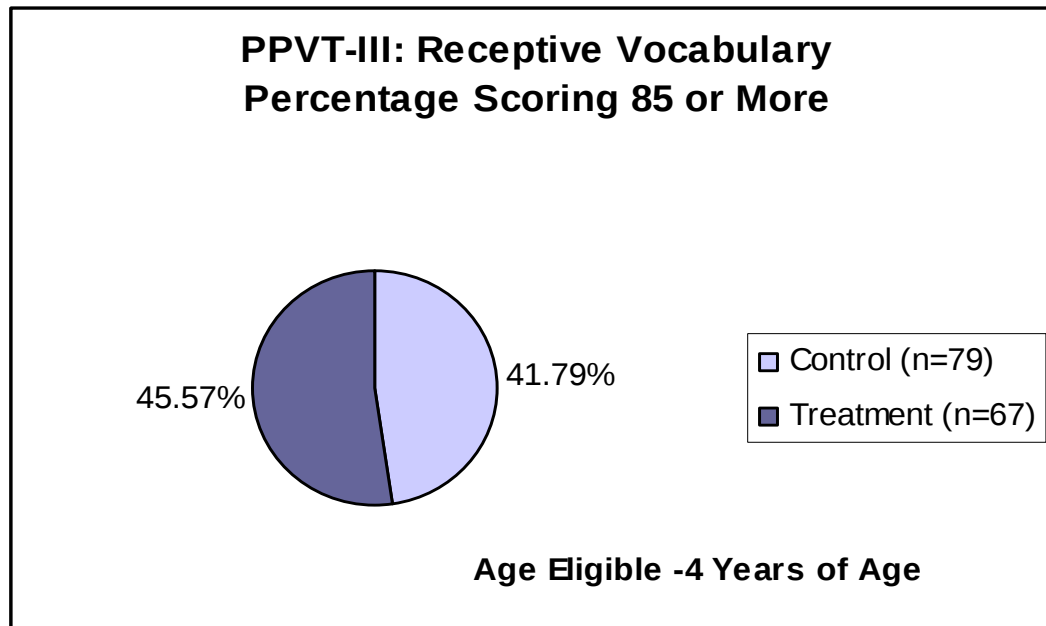


Figure 9. Impacts on Children's Receptive Vo: Post-Course 1

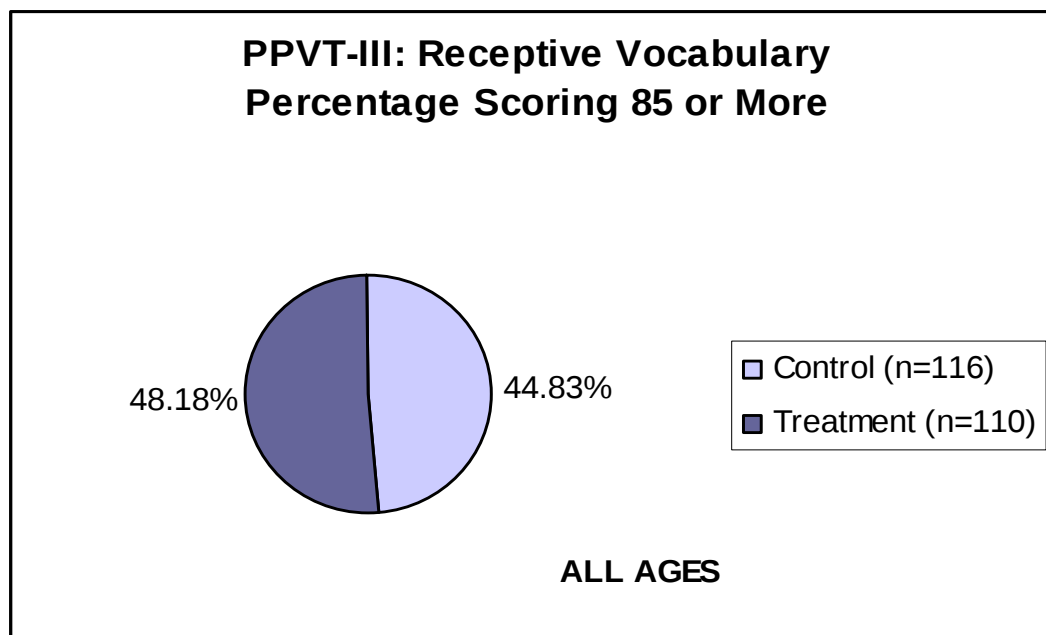
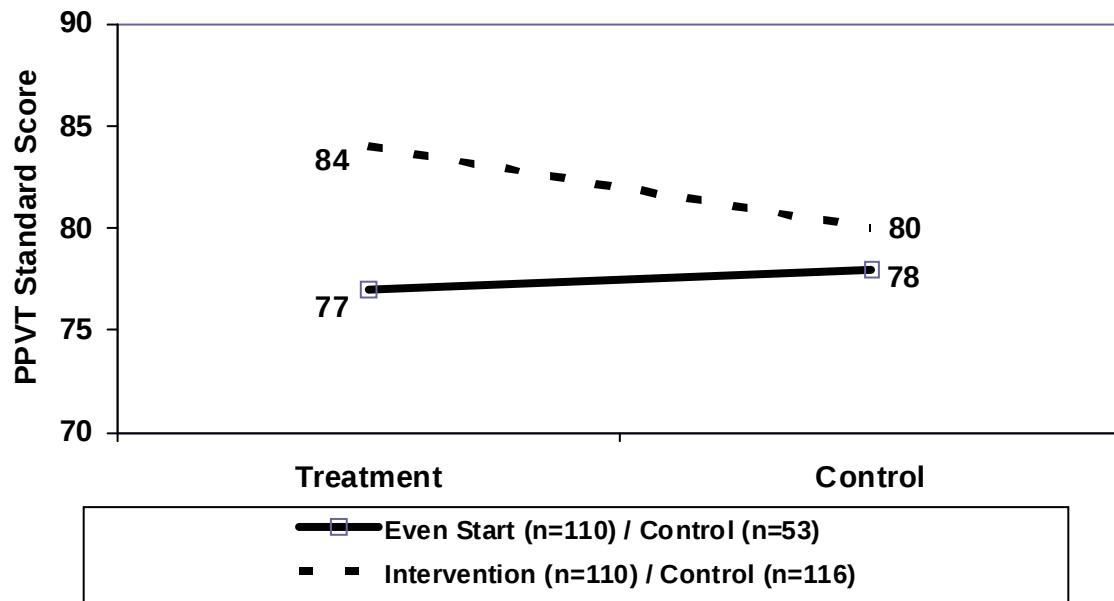


Figure 10. Intervention PPVT Scores as Compared to Even Start



Source: Chart modified from the one presented in the Third National Even Start Evaluation: Follow-Up Findings From the Experimental Design Study (Ricciuti et al., 2004).

APA Reference Style Examples

Sample Citation: Journal Article

Hypericum Depression Trial Study Group. (2002). Effect of Hypericum perforatum (St John's Wort) in major depressive disorder: A randomized controlled trial. *JAMA*, 287, 1807–1814.

Sample Citation: Newsletter/Newspaper Article

Brown, L. S. (1993, Spring). My research with oranges. *The Psychology Department Newsletter*, 3, 2.

Sample Citation: Book

American Psychiatric Association. (1990). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Booth, W. C., Colomb, G. G., & Williams, J. M. (1995). *The craft of research*. Chicago: University of Chicago Press.

Sample Citation: Chapter or Section in a Book

Stephan, W. G. (1985). Intergroup relations. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., Vol. 2, pp. 599–658). New York: Random House.

Sample Citation: Web Page

Dewey, R. A. (2004). *APA Style Resources by Russ Dewey*. Retrieved September 8, 2004, from <http://www.psywww.com/resource/apacrib.htm>